A Pilot Study for BERT Language Modelling and Morphological Analysis for Ancient and Medieval Greek

Pranaydeep Singh, Gorik Rutten and Els Lefever
LT3, Language and Translation Technology Team
Department of Translation, Interpreting and Communication – Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
firstname.lastname@ugent.be

Abstract

This paper presents a pilot study to automatic linguistic preprocessing of Ancient and Byzantine Greek, and morphological analysis more specifically. To this end, a novel subword-based BERT language model was trained on the basis of a varied corpus of Modern, Ancient and Post-classical Greek texts. Consequently, the obtained BERT embeddings were incorporated to train a fine-grained Part-of-Speech tagger for Ancient and Byzantine Greek. In addition, a corpus of Greek Epigrams was manually annotated and the resulting gold standard was used to evaluate the performance of the morphological analyser on Byzantine Greek. The experimental results show very good perplexity scores (4.9) for the BERT language model and state-of-the-art performance for the fine-grained Part-of-Speech tagger for in-domain data (treebanks containing a mixture of Classical and Medieval Greek), as well as for the newly created Byzantine Greek gold standard data set. The language models and associated code are made available for use at https://github.com/pranaydeeps/Ancient-Greek-BERT

1 Introduction

During the last decades, large collections of digital texts have become available for Ancient Greek and Latin. As a result, classicists are becoming more and more interested to apply Natural Language Processing (NLP) techniques to extract information from these texts in an automatic and structured way. Although there has been a boost in NLP research for Greek and Latin thanks to the introduction of the Classical Language Tool Kit[1] and the development of Dependency Treebanks, such as the The Ancient Greek and Latin Dependency Treebank (AGLDT)[2], there is still a lack of NLP tools that perform well on different historical varieties, such as for instance Byzantine Greek[3].

The presented research is to be situated in the overarching DBBE project[4], where Byzantine epigrams are studied as nodes between textual transmission and cultural and linguistic developments (Bernard and Demoen, 2019). This multidisciplinary project aims to reveal the connections between linguistic patterns and text-historical developments in a corpus of metrical paratexts in Byzantine Greek manuscripts and will develop new digital tools designed to perform linguistic analysis and to detect patterns and variations in this fragmented corpus of Byzantine text.

The linguistic analysis of Ancient (and Medieval) Greek has some challenges. The free word order is expected to create difficulties for automatic linguistic preprocessing, but both Dik and Whaling (2008) and Keersmaekers (2019) conclude that the word order does not cause specific problems for the task of Part-of-Speech (PoS) tagging. Considering we are interested in Byzantine Greek in particular, it is also important to mention there are major differences between Classical Greek and Byzantine Greek, which can be observed on the phonetic, phonological, morphological, syntactical and lexical level. In some cases, Byzantine Greek is reminiscent of Modern Greek, rather than Ancient Greek. As a consequence, many of the systems that are widely used for the analysis of Ancient Greek are expected to struggle with Byzantine Greek. The very popular morphological analysis tool Morpheus (cf. infra) in particular, is expected to not operate well for Medieval Greek.

The remainder of this paper is organized as follows. In Section 2, we briefly discuss some of the relevant research related to linguistic preprocessing of Ancient Greek. In Section 3, we describe the

[1] http://cltk.org/
[2] https://perseusdl.github.io/treebank_data/
[3] In this paper, *Classical* and *Medieval* Greek are used interchangeably with *Ancient* and *Byzantine* Greek respectively.
[4] https://www.dbbe.ugent.be/

Proceedings of LaTeCH-CLfL 2021, pages 128–137
Punta Cana, Dominican Republic (Online), November 11, 2021.

data set used to train and evaluate the morphological analyser and provide details on the creation of a novel Byzantine gold standard data set. Section 4 gives an overview of the BERT language model creation for Ancient and medieval Greek, whereas Section 5 describes the fine-tuning of the resulting model on the task of morphological analysis. In Section 6, we end with concluding remarks as well as indications for future research.

2 Related Research

Previous research has resulted in various tools and resources for the linguistic analysis of Ancient Greek.

The first systems to perform NLP tasks on Ancient Greek relied on morphological analysis tools. In the early 1970s, Packard (1973) already wrote a program to analyze Greek words with a 95% accuracy. Great strides forward were achieved in the 1990s, when the morphological Analysis tool Morpheus was developed in the framework of the PERSEUS project (Kent, 1991). From this point onward, Morpheus has been the standard tool for Ancient Greek morphological analysis. In his paper on Generating and Parsing Classical Greek, Crane (1991) describes how Morpheus operates by combining information from three databases: one database containing stems, a second one containing possible inflections and a third one containing irregular forms. Although this system is accurate in optimal circumstances, it has two distinct weaknesses: (1) all possible morphological analyses are provided, but Morpheus does not decide which analysis is correct in casu, and (2) the rule-based nature does not deal well with unknown words or non-classical Greek forms. The system can deal with some historical and dialectal variation to a certain extent, but Byzantine Greek is not within the scope of Morpheus. A possible remedy consists in manually adding words to Morpheus' vocabulary as shown by Keersmaekers (2019). This approach, however, does not scale very well nor would it remedy the issue of the conjugation and declension of existing words changing. Nevertheless, we can observe that in many of the approaches presented for Ancient Greek, the morphological information provided by Morpheus is fundamental in the analysis.

2.1 Part-of-Speech Tagging

Part-of-Speech tagging is the NLP task of automatically assigning a morphological label to each token in a text, and traditionally refers to assigning a coarse-grained grammatical category, viz. a *part of speech* (such as noun, adjective, verb) to every token. For the analysis of Ancient Greek, however, Part-of-Speech tagging usually refers to a more fine-grained, full morphological analysis, i.e., the part of speech and morphological features such as gender, person, number, mood or tense. Therefore, we will also use the term *Part-of-Speech tagging* and (full) *morphological analysis* interchangeably in this paper.

Treetagger (Schmid, 1994), a probabilistic tagging method employing decision trees, was one of the first Part-of-Speech taggers used to analyse Ancient Greek. Dik and Whaling (2008) enhanced the tool with a lexicon from Morpheus. The efficacy of Treetagger is proven by its continued use: even recently the tagger has been used to achieve state-of-the-art results on the Diorisis Ancient Greek Corpus (Mcgillivray and Vatri, 2018).

Celano, Crane and Majidi (2016) compared five different PoS-taggers for Ancient Greek: Mate[5], Hunpos[6], RFTagger[7], OpenNLP[8] and NLTK unigram tagger[9]. They perform their comparison using a 10-fold cross-validation test on the Ancient Greek Dependency Treebank (AGDT). In this comparison, Morpheus was applied to provide all possible morphological analyses for each word, from which the correct one was manually chosen in order to construct the training (and test) data. Mate achieved the best results (88% accuracy), followed by the other taggers. This relatively low accuracy can be attributed to the very fine-grained nature of the analyses. It is important to keep the standard, classical nature of the texts in mind here when observing the tagging results.

Keersmaekers (2019) offers a second look at the different available part-of-speech taggers. He compares RFTagger, MarMoT[10] and Mate. The results obtained from this study differ from the aforementioned study by Celano et al. (2016). Mate was outperformed by both RFTagger and MarMoT on

[5]https://www.ims.uni-stuttgart.de/en/research/resources/tools/matetools/

[6]https://code.google.com/archive/p/hunpos/

[7]https://www.cis.uni-muenchen.de/ schmid/tools/RFTagger/

[8]https://opennlp.apache.org/

[9]https://www.nltk.org/book/ch05.html

[10]http://cistern.cis.lmu.de/marmot/

a papyrus test corpus. With a reported accuracy of 94.7%, RFTagger and MarMoT both exceed expectations. It is worth mentioning that the addition of a lexicon from Morpheus increased the accuracy of RFTagger by 2.4%. Keersmaekers attributes these divergent results to the difference between tag sets. The relatively worse performance of Mate was attributed to the following factors: (1) the tagging model being unsuitable for Greek, (2) the smaller amount of training data, and (3) the joint parsing model of Mate could be detrimental due to low parsing accuracy. Just like the Byzantine data that we want to analyze for this project, the test data used for this comparison also diverges from classical Ancient Greek, albeit in a different way. At the one hand, the Greek found in the papyrus corpus is closer to Ancient Greek on the morphological level. On the other hand, different constructions can be expected to be encountered on the syntactical level. All in all, the Greek found in the papyrus data can more accurately be analyzed by Morpheus, which will constitute a fundamental difference in possible viable approaches.

The problem caused by the incompatibility of Morpheus with Byzantine is a major concern for our research. Since the final goal is to develop a preprocessing pipeline that will work for the very varied Byzantine DBBE corpus, which contains a lot of out-of-vocabulary words, the use of Morpheus is preferably avoided. As mentioned already, it would be possible to manually add words to partially alleviate this issue, but this is not a viable solution due to the size of the DBBE.

More recently, researchers have started creating Part-of-Speech taggers not incorporating information from Morpheus. Helmut Schmid's most recent creation, the RNNTagger, was of particular interest due to an increased tagging accuracy on the AGDT (Schmid, 2019). The reported results, viz. accuracy scores of 91%, are very impressive considering this system is no longer dependent on separate output from Morpheus. Additionally, if RNNTagger would be trained on sufficient Byzantine Greek data, it is likely to perform similarly, and would be the first system to perform Part-of-Speech tagging on Byzantine Greek with such high accuracy. Unfortunately, no such data is publicly available as of yet.

Finally, some other recent developments can also be mentioned here. Stanford university's Stanza package (Qi et al. 2020) contains two different PoS-taggers, each one trained on a different database, viz. the AGLDT and PROIEL treebanks. A shallow evaluation we performed for these taggers showed that RNNTagger outperformed both of them.

2.2 Language Models for Ancient Greek

Recent state-of-the-art PoS-taggers (see for instance Heinzerling and Strube (2019)) often integrate information from a BERT language model. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) is based on Transformers, a deep learning model where every output node is linked to every input node, whereas the weights between them are dynamically computed during training. As opposed to directional models, which read the input sentence from left-to-right (or right-to-left), the Transformer model is "bidirectional", i.e. it reads the entire sentence at once. As a result, the model learns the representation of a word based on both its left and right context.

Usually, such a BERT neural language model is pre-trained on a huge data set for a particular target language, and then subsequently, the neural network is used as the basis to fine-tune for a specific NLP taks, such as PoS-tagging.

To the best of our knowledge, only Brennan Nicholson has trained a BERT model for Ancient Greek[11], which makes use of character-based embeddings. This model can, however, not easily be fine-tuned to perform tasks other than masked word prediction, since it's a character-based model unlike the standard sub-word model that we strive to train. Therefore, we present in this research a new subword-based BERT model that can be fine-tuned to perform various tasks like morphological analysis for Ancient and Medieval Greek. More details on the creation of this language model can be found in Section 4.

3 Training and Evaluation Data Sets for Morphological Analysis

To train (and evaluate) our novel morphological analyser, we extracted the relevant information from three different treebanks (Section 3.1). To evaluate the resulting model on Byzantine Greek, a novel gold standard data set was created (Section 3.2).

[11] https://github.com/brennannicholson/ancient-greek-char-bert

3.1 Training Data Morphological Analyser

In order to train and validate a morphological analyser for Ancient and Medieval Greek, we used three existing treebanks, viz. the PROIEL treebank[12], the Ancient Greek Dependency treebank[13] and the treebank created by Vanessa Gorman[14].

The **PROIEL treebank** (Haug and Jøhndal, 2008) consists of three different texts: (1) The Greek New Testament, (2) Herodotus' Histories, and (3) Sphrantzes' chronicles. As such, part of the treebank is Byzantine Greek, while the other part stems from the classical period.

The **Ancient Greek Dependency Treebank** (Bamman and Crane, 2011) includes exclusively Classical Greek data. It currently contains 557,922 tokens of various different authors, genres and dialects, ranging from Homer's Epics to various of Sophocles' tragedies, different dialects and periods. The most recent text present in this database is Athenaeus' Deipnosophists.

Gorman's Treebank (Gorman, 2020) aims to collect representative texts from Ancient Greek prose authors. This monumental work annotated by the author contains over 550,000 tokens of Classical Greek Prose, and no Byzantine data.

In conclusion, the main part of the data used to train the morphological analyser is written in Classical Greek, but the presence of a reasonable amount of Byzantine Greek should not be understated. This Byzantine fraction will be likely to help the model in analyses of Byzantine Greek.

To create the training data set for the development of the fine-grained PoS-model, we extracted all tokens and corresponding Part-of-Speech tags from all three treebanks, that had already been converted to a uniform format by Keersmaekers (2020). It is important to mention, though, that the Part-of-Speech tags present in the treebanks have been further refined by Keersmaekers (2020), who for instance makes a distinction between coordinate and subordinate conjunctions, which is not present in the original treebank annotations.

As mentioned before, the Part-of-Speech tags used in this research contain a full morphological analysis as well. The predicted labels consist of nine parts: coarse-grained part-of-speech category, person, number, tense, mood, voice, gender, case, and degree. Elements that are not relevant for a given token are represented by '-'. The label *v2spme—*, for instance, stands for a verb - in the second person - singular - present tense - imperative - mediopassive. This fine-grained morphological information leads to a high number of possible labels, resulting in 558 different class labels for the Gorman Treebank, 483 for the PROIEL Treebank and 599 for AGDT. Needless to say that predicting these fine-grained morphological labels makes the present machine learning task far more challenging than the more traditional coarse-grained Part-of-Speech tagging task.

3.2 Gold Standard Creation for Byzantine Greek

In order to create a gold standard for evaluating the morphological analyser for medieval Greek, we manually annotated part of the existing Database of Byzantine Book Epigrams (henceforth DBBE). This database, made and maintained by Ghent University, consists of Greek book epigrams (poems in and on books) up to the fifteenth century. Two kinds of textual material are distinguished and defined as follows by the DBBE team:

Occurrences: "Book epigrams exactly as they occur in one specific manuscript. The data collected here is largely the result of careful manuscript consultation, either in situ or based on (digital) reproductions. The remainder is compiled from descriptive catalogues and other relevant publications. Individual verses found in multiple occurrences are linked together by means of dedicated Verse variants pages."

Types: "Book epigrams independently of how exactly they occur in the manuscripts, often, yet not always, regrouping several occurrences that have an identical or at least very similar text. If available, the text of a type is drawn from a critical edition. If not, it is a normalised version of a single representative occurrence." For the presented pilot study, we focused on the "Types" data. In future research, we will also annotate part of the more irregular "Occurrences" data.

To create the gold standard, we selected about 9000 tokens (which amounts to 6,5% of the complete "Types" database). To speed up the manual validation process, we bootstrapped the PoS information from the output of the RNN-tagger (Schmid, 2019), that performed reasonably well on this data.

[12] https://proiel.github.io/

[13] https://perseusdl.github.io/treebank$_data$/

[14] https://github.com/perseids-publications/gorman-trees

Ancient Greek BERT trained on text from treebanks with random initialisation	18.0
+ Initialization from bert-greek-uncased instead of random **Mixed Greek BERT**	9.8
+ Data from Perseus Digital Library and First1K Greek Database **Expanded Ancient Greek BERT**	4.9

Table 1: Content and perplexity scores of the three flavours of BERT language models trained for Ancient Greek.

Four peculiarities of RNNTagger, however, came to light during the manual correction of the output:

1. Incomplete pronoun PoS-tags: in many cases (62/96 in a manually analysed gold standard sample), RNNTagger did not correctly tag pronouns with the correct person. This is particularly strange as it managed to predict the lemma as the first person singular pronoun in all of these cases, but did not indicate the first person in the provided Part-of-Speech tag.

2. Consistency problems: the same word is analysed in different ways in different loci, even if only one analysis is possible for that specific word.

3. Lemmatization and Part-of-Speech tagging are not corresponding: even when words are analysed as a singular nominative, meaning that the analysed word would in all cases correspond with the lemma, the predicted lemma surprisingly suggests a different word for about one in five tokens.

4. Generally poor predicted lemmata: continuing on the previous remark, it can be noted that the predicted lemmata are generally imprecise (roughly obtaining a 50% accuracy). Especially substantives, adjectives and verbs suffer from this problem, whereas pronouns seem less problematic. As this study focuses on Part-of-Speech tagging only, this is not a major concern, but it might be relevant for further research.

After the output from RNNTagger was manually corrected, an analysis from our own Part-of-Speech tagger output (as discussed in chapter 5) allowed us to correct remaining errors in the gold standard labeling.

In future research, we plan to further annotate the complete DBBE data set, which will make the resulting gold standard also suited to fine-tune a system for PoS-tagging using this Byzantine corpus.

4 Language Modeling with BERT

A shallow linguistic analysis showed that the actual nature of the "Types" diverges from the expected Byzantine Greek. Many of the byzantine irregularities have been normalized, and the text of the Types contains a lot of classical structures. Consequently, this will cause many of the existing systems for Ancient Greek to work remarkably well. Therefore, we decided to train our final BERT language model on a combination of Ancient Greek and Modern Greek data, as the Types are definitely more similar to Ancient Greek than to Modern Greek.

4.1 Ancient Greek Data Used to Train the Model

BERT models are very data greedy. Consequently, we did not only rely on the treebank databases to train the language model, but also used the (much larger) full text databases available for Ancient Greek. The text databases used were the following: (1) The (Ancient Greek part of the) Perseus Digital Library and (2) The First1KGreek Project Database.

The Perseus Digital Library currently contains 13,407,448 words of Classical Greek texts. The First1KGreek collects one edition of every Greek work composed between Homer (9th century BC) and 250 AD, that does not occur in the Perseus Digital Library. The combination of these two databases results in a relatively complete representation of Classical Greek.

4.2 BERT Model for Ancient Greek

In this research, we opted to first train a generic BERT language model for Ancient and Byzantine Greek, and then in a second step fine-tune the model for the task of morphological analysis.

A first version of the BERT language model was trained primarily on Ancient and Byzantine Greek data, considering the affinity of the "Types" and An-

cient Greek data. All available Greek texts present in the three treebanks were used to train the model. The model was trained with following conditions: 15% of the words are replaced with [MASK] tokens, the maximum sequence length per sentence was limited to 512 sub-words and we experimented with two random initialisations. The model optimizes for the Masked Language Modelling objective which is formulated as:

$$-\sum_{w=1}^{V} y_{o,w} log(p_{o,w}) \quad (1)$$

where V is the size of the vocabulary, $y_{o,w}$ is a binary indicator, 0 or 1, for the word w being the correct replacement for [MASK], while $p_{o,w}$ is the probability predicted by the Transformer of the word w being a good replacement.

In order to evaluate the various iterations of our language model, we use perplexity as a metric, showing how uncertain any given system is. More formally, Perplexity is the inverse probability of the test set, normalized by the number of words as shown in the formula below:

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1, w_2 ....., w_n)}} \quad (2)$$

The perplexity for this first iteration of our BERT model trained on the available Ancient Greek corpora was 18.0, whereas perplexity of a good language model is expected to be around 10. However, the texts in question did not have a massive vocabulary comparable to, for example, English. Therefore, theoretically the perplexity should be well under 10.

In order to improve (i.e. decrease) the perplexity of our model, we tried to leverage the existing resources for Modern Greek. We started from the publicly available BERT-Greek-uncased (Koutsikakis et al., 2020) pre-trained model. This BERT model for Modern Greek was trained on various modern greek resources like the Greek Wikipedia. Even though this model initially did not perform well on Ancient Greek (perplexity >20), the pre-training can serve as a better starting point than random initialisation, helping it understand the Greek characters and some common parts of the vocabulary. After training this model on the Ancient Greek corpus using MLM, performance was significantly increased and the perplexity dropped to 9.8 on the held-out test set. From now on, we will refer to this model as the **Mixed Greek BERT model**.

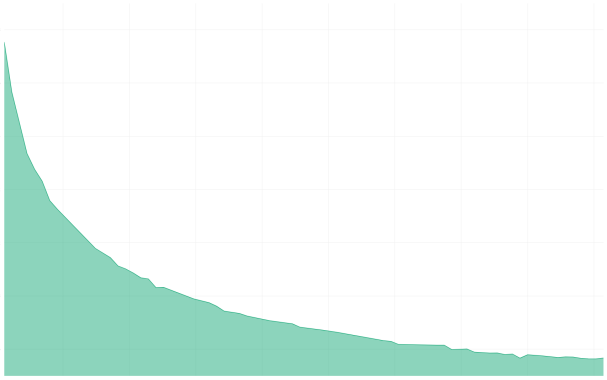

Figure 1: The convergence of loss for the Expanded Ancient Greek BERT model on the held out test set.

However, since we hypothesize that the theoretical perplexity could be much lower than that of a standard BERT model, we wanted to further explore the upper bound of an Ancient Greek BERT. We built a third and final model starting from the pre-trained Modern Greek model, but also including the two large text collections described in Section 4.1, the Greek part of the Perseus Digital library and the First1KGreek database, in addition to the treebank data, overall increasing the original training data by around 300 percent. We also used special lower-casing for Greek, and de-accentuated the text as a pre-processing step. This new model considerably outperforms all the previous iterations and results in a perplexity of 4.9 on a held-out test set. The validation loss convergence as a function of time can be seen in Figure 1. This model is henceforth referred to as the **Expanded Ancient Greek BERT** model. This pre-trained language model has been made publicly available, together with some sample code showing how to use it[15]. Table 1 summarizes the content and perplexity scores of the three flavours of BERT language models we trained for Ancient and Byzantine Greek.

5 Fine-tuning of the BERT model for Morphological Analysis

In a last step, we incorporated the novel BERT language model for Ancient and Byzantine Greek into our morphological analysis system.

[15] https://github.com/pranaydeeps/Ancient-Greek-BERT

5.1 Training of the Morphological Analyser

As explained in Section 3.1, we extracted all tokens and corresponding fine-grained PoS-tags from the three treebanks. We use the contextual token embeddings from the Expanded Ancient Greek BERT model described in Section 4.2, and stack them with randomly initialized Character Embeddings (Lample et al., 2016), which are then processed by a standard Bi-directional Long Short Term Memory (LSTM) encoder and a Conditional Random Field (CRF) decoder, commonly used in sequential tagging tasks. The Expanded Ancient Greek BERT embeddings are the only frozen part of the model, while all the other fragments are trained. We use the FLAIR framework (Akbik et al., 2019) for this set of experiments. While the Bi-LSTM CRF architecture is a staple of many successful sequential taggers, we elect to stack our BERT embeddings with character embeddings that have found extensive use in models for languages with high morphological diversity (Vylomova et al., 2017). We use a hidden size of 256 for our LSTM, and initialize with a starting learning rate of 0.1, which is linearly decreased.

5.2 Evaluation of the Morphological Analyser

For the evaluation of the tagger it is important to keep in mind that we are analyzing Byzantine Greek with models trained mainly on Ancient Greek. The tagger's accuracy will therefore inevitably be lower. Second, inter-annotator agreement on Greek PoS-tagging is most likely a lot lower than for other languages. In various cases, different analyses can be considered correct by experts. Participles, for example, form a middle ground between verbs on the one hand and substantives and adjectives on the other hand. In some cases, a participle that was repeatedly used in a certain form is accepted as a noun or adjective by linguists. This status, however, depends on convention. As such, it is very difficult for a system to detect whether a participle is considered a noun, adjective of verb in many occasions. Even for humans, this issue can be hard to resolve, arguments can sometimes be made for different analyses. For this pilot study, we only constructed a modest gold standard of 9000 tokens. In future research, we intend to construct detailed annotation guidelines and perform inter-annotator experiments to further expand the current gold standard, based on the insights gained during this pilot study.

5.2.1 Validation on the in-domain data

The performance of our trained model obtains state-of-the-art results on the very fine-grained PoS scheme incorporating a full morphological analysis as applied in the treebanks. Table 2 shows the accuracy scores for the different treebank validation sets. These validation sets have been randomly selected from the treebanks data, and have not been used for training the morphological analysis model. The scores presented are the result of a four-fold cross-validation evaluation.

Treebank	Accuracy
AGDT	88.64%
PROIEL	92.87%
GORMAN	91.85%

Table 2: Accuracy of our morphological analysis tagger on in-domain data, viz. validation sets of the three treebanks used for training.

When we compare this to reported scores for the various treebanks, the newly trained PoS-tagger obtains very competitive results. The study of Celano et al. (2016) comparing five taggers for Ancient Greek reports the best score for Mate on the AGDT, incorporating information from Morpheus, with an accuracy of 88%. It is important to note, however, that the scores are not directly comparable, as they are not obtained on exactly the same data partitions.

When we only consider the coarse-grained PoS-tags, viz. the syntactic categories (e.g. *noun, verb*), the results are comparable to state-of-the-art PoS-taggers for other languages. Table 3 gives an overview of the accuracy scores when only taking into account the coarse-grained PoS-tags.

Treebank	Accuracy
AGDT	90.28%
PROIEL	97.40%
GORMAN	95.71%

Table 3: Accuracy for the coarse-grained PoS-tags on in-domain data.

5.2.2 Evaluation on the held-out test data

The gold standard that was created allows to evaluate the performance of existing PoS-taggers on Byzantine Greek data. For this pilot study, we compare RNNTagger with our newly created PoS-tagger. As is illustrated by Table 4, our PoS-tagger

incorporating the new BERT language model obtains very promising results, with an accuracy of 86.88% for the full morphological analysis. When we only consider the coarse-grained PoS-tags, the accuracy increases to 92.97%.

Fine-grained PoS Tagger	Accuracy
RNNTagger	76.97%
BERT model	86.88%

Table 4: Accuracy of the two fine-grained PoS-taggers on the Byzantine gold standard data set.

We also performed a qualitative analyse of both taggers' output. The tags predicted by RNNTagger structurally subtly differ from the tags in our gold standard. This leads to reduced accuracy. Our training data, for instance, distinguishes between coordinating and subordinating conjunctions whereas RNNTagger does not make this distinction. On the other hand, in cases where several interpretations were possible, RNNTagger's interpretation is more likely to be adopted as manual correction started from RNNTagger's output.

To offer some insights into the frequency of the detected errors, a small test sample containing 523 fine-grained PoS-tags of RNNTagger and our own morphological analyser (discussed in section 5) was examined and compared in detail. RNNTagger mainly struggles with the following categories:

- Providing full tags for personal pronouns: the person is often omitted (in 8/8 cases in our small test sample).
- Consistently analysing words with only one possible correct analysis (2 occurrences of this phenomenon in the small test sample).
- Recognizing Byzantine names and vocatives in general (8 errors due to incorrectly analysed Byzantine names).
- Detecting '·' as a punctuation mark (the Greek ':', not to be confused with '.'). This is incorrectly analysed as a part of the previous word in about 95% of cases.

On our gold standard, RNNTagger was 76.97% accurate. Considering the nature of the Byzantine text and the slight mismatch between the annotations of the gold standard and RNNTagger, this accuracy is in line with the expectations. Pronouns not being recognized as such, and the wrongly tagged proper names, caused a significant decrease for the accuracy of the tagger.

Finally, our own model achieved an accuracy of 86.88%. The tagger's most notable errors observed were the following:

- Iota subscriptum (a iota in subscript) was often not recognized by the model, leading to incorrect analyses.
- Parentheses and other symbols indicating uncertainty in the text caused the model to not make sensible predictions for the word in question. Even an apostrophe would cause wrong PoS-analyses.
- The perfect tense was sometimes not recognized.
- Proper names that are indeclinable are analysed as a noun without gender, case and number. The gold standard does, however, include this more detailed information, causing all indeclinable names to be evaluated as incorrect.
- As was the case with RNNTagger, '·' is never analysed as a punctuation mark, but as part of the previous word instead.

It is important to remark that comparing RNNTagger with our newly developed morphological analyser is a difficult exercise. The output of both taggers diverges in a different way from the gold standard. In 26 cases in our small test sample of 523 words, an error resulted from an analysis that was in line with how the respective tagger was trained, but was evaluated as incorrect because the predicted label was not corresponding with our gold standard labeling.

To conclude, our new morphological analysis model resolves most of the issues RNNTagger struggled with: Byzantine names and vocatives were almost always correctly analysed, full pronoun tags were provided, and results were consistent. Punctuation, however, appears to be challenging for the new model as well.

6 Conclusions

In this pilot study, we first present a new BERT-based language model for Ancient and Byzantine Greek, the so-called Expanded Ancient Greek BERT model, obtaining very good perplexity results. The model starts from a pre-trained Modern Greek language model, to which Ancient and

Byzantine Greek data (Perseus Digital library, First1KGreek database and AGDT, Proiel and Gorman treebank data) is added. The performance of the language model was further improved by using special lowercasing and de-accentuation for Greek.

Second, we fine-tuned a novel morphological analysis model for Ancient and Byzantine Greek, that does not rely on separate morphological input from Morpheus, as is the case for many existing taggers. As a result, this model is more flexible to analyse irregular forms in Ancient Greek and will perform better at analysing Greek that does not originate from the classical period. The model achieves state-of-the-art performance on a validation set extracted from the PROIEL, AGDT and Gorman treebank data sets. The model also obtains very promising results on a newly created gold standard consisting of Byzantine Greek epigrams.

In future research, we will build on the insights from this pilot study to improve the performance of the fine-grained Part-of-Speech tagger for Byzantine Greek. A first line of research will consist of establishing detailed annotation guidelines and conducting inter-annotator experiments to further improve and expand the gold standard data set. Once a considerable amount of the DBBE will be annotated, this data can be used to further fine-tune the morphological analyser for Byzantine Greek data.

Another venue for future research will consist of adding more diverse Greek data, e.g. Byzatine data from the DBBE, to update the language model, and evaluate whether this impacts the performance of the tagger on Byzantine Greek data.

Finally, we will also experiment with a cascaded approach predicting the various parts of the morphological analysis consecutively, starting with the coarse-grained part-of-speech category (e.g., *noun, verb*). This allows to separate the "classical" part-of-speech from the fine-grained morphological analysis, providing more evidence for each part of the analysis. This approach might, however, introduce error percolation from the various consecutive steps.

To conclude, we are confident that the newly developed BERT-based language model will be a valuable contribution to NLP for Ancient and Medieval Greek, as it can be fine-tuned for a variety of downstream tasks, such as linguistic analysis.

Acknowledgements

Special thanks go out to Dr. Alek Keersmaekers, who provided us with uniform versions of the three treebanks. We also thank Dr. Ilse De Vos, Marthe Nemegeer and Noor Vanhoe from the DBBE team for their help with the creation of the gold standard evaluation set.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

D. Bamman and G. Crane. 2011. The ancient greek and latin dependency treebanks. pages 79–98. Springer.

F. Bernard and K. Demoen. 2019. Book epigrams. a companion to byzantine poetry. volume 4, pages 404–429. Brill.

G. Celano, G. Crane, and S. Majidi. 2016. Part of speech tagging for ancient greek. *Open Linguistics*, 2:393–399.

G. Crane. 1991. Generating and parsing classical greek. literary and linguistic computing. *Literary and Linguistic Computing*, 6:243–245.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

H. Dik and R. Whaling. 2008. Bootstrapping classical greek morphology. In *Proceedings of Digital Humanities 2008*, pages 105–106, Oulu.

V.B. Gorman. 2020. Dependency treebanks of ancient greek prose. *Journal of Open Humanities Data*, 6.

D. T. Haug and M. Jøhndal. 2008. Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34.

Benjamin Heinzerling and Michael Strube. 2019. Sequence tagging with contextual and non-contextual subword representations: A multilingual evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 273–291, Florence, Italy. Association for Computational Linguistics.

A. Keersmaekers. 2019. Creating a richly annotated corpus of papyrological greek: The possibilities of natural language processing approaches to a highly inflected historical language. *Digital Scholarship in the Humanities*, 35:67–82.

A. Keersmaekers. 2020. A computational approach to the greek papyri: Developing a corpus to study variation and change in the post-classical greek complementation system.

A. Kent. 1991. *Encyclopedia of Library and Information Science: Volume 48 - Supplement 11: Automated Archival Systems to University-Based Technology Transfer and 2000: Explanation: Example, and Expectations*. CRC Press.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. GREEK-BERT: the greeks visiting sesame street. In *SETN 2020: 11th Hellenic Conference on Artificial Intelligence, Athens, Greece, September 2-4, 2020*, pages 110–117. ACM.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

B. Mcgillivray and A. Vatri. 2018. The diorisis ancient greek corpus. *Research Data Journal for the Humanities and Social Sciences*, 3:55–65.

D.W. Packard. 1973. Computer-assisted morphological analysis of ancient greek. In *COLING 1973 Volume 2: Computational And Mathematical Linguistics: Proceedings of the International Conference on Computational Linguistics*.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

H. Schmid. 2019. Deep learning-based morphological taggers and lemmatizers for annotating historical texts. In *DATeCH2019: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, pages 133–137.

Ekaterina Vylomova, Trevor Cohn, Xuanli He, and Gholamreza Haffari. 2017. Word representation models for morphologically rich languages in neural machine translation. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 103–108, Copenhagen, Denmark. Association for Computational Linguistics.